Data Mining for Cardiovascular Disease Prediction

Islam Elgarhy
Electrical and Computer Engineering Department, College of Engineering
Tennessee Technological University
Cookeville, TN, USA
iaelgarhy42@tntec.edu

***Abstract*— Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels and are a major cause of disability and early death worldwide. For example, in the USA, one person dies every 36 seconds due to CVDs. In addition, it affects national income due to the cost of health care services, medicines, and lost productivity due to death. It's important to notify the individual at higher risk of developing CVD to prevent early deaths. Most often it's challenging for medical practitioners to predict cardiovascular disease as it requires experience and knowledge. The advances in the field of computational intelligence, together with the massive amount of data produced every day in clinical settings, have made it possible to create recognition systems capable of predicting whether an individual has a CVD. Support Vector Machine (SVM), and Convolutional Neural Network (CNN) will be used to train on the Kaggle dataset of CVD cases, which includes 70000 registers of patients and 12 attributes divided into three types (Objective, Examination, and Subjective) considered relevant for identifying the disease. A feature weight is used to select which features are more useful in the training process in order to achieve a better accuracy.**

Keywords—Cardiovascular Diseases (CVDs), Data Mining, Support Vector Machine (SVM), Convolutional Neural Network (CNN)

I. INTRODUCTION

Recently, cardiovascular diseases (CVDs) are the leading cause of early death all over the world irrespective of gender. In 2019, approximately 17.9 million people died from CVDs, representing 32% of all global deaths. Over three-quarters of CVDs deaths take place in low- and middle- income countries. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, harmful use of alcohol, unhealthy diet, obesity, and physical inactivity [1].

It is important to detect CVDs as early as possible to prevent early deaths and begin the management with counseling and medicines which are described by doctors. In contrast, most often it's challenging for doctors to predict CVDs as it requires experience and knowledge which is a complex task to accomplish [1].

There is an affluence of data obtained in the healthcare industry. Data is useful for making effective conclusions using their hidden information, so applying data mining algorithms on this type of data plays a significant role in the prediction and diagnosis of diseases.

Data mining is the process of discovering useful information in large data repositories automatically. Some data mining algorithms focus on the predictive process where a prediction of the value of a particular attribute is based on the values of other attributes, The predicted attribute is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables [2]. Due to the negative effect of poor data quality on data mining efforts, several data issues need to be considered and require some data preprocessing techniques for successful data mining algorithms.

Support Vector Machine (SVM), and Convolutional Neural Network (CNN) are two data mining algorithms, which can be used to predict CVDs [2]. SVM is widely used in both classification and regression processes. SVM is a binary classifier; One-to-One and One-to-Many are two approaches to implement multiclass. CNN is a class of artificial neural networks that have been designed to learn spatial hierarchies of features automatically and adaptively through backpropagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers [2].

The rest of this paper was organized as follows: Section II discusses a few related works. The dataset is described in Section III. Methodology and proposed model are discussed in Section IV. Experimental results are given in Section V. Section VI shows the conclusion and future work.

II. RELATED WORK

There is a lot of research that has focused on CVDs prediction using different data mining techniques on different datasets [3].

In 2021, Barbara Martins [4] proposed five classifiers namely DT, Optimized DT, RI, RF, and DL. The models were mainly developed using the RapidMiner software with the assist of the WEKA tool and were analyzed based on accuracy, precision, sensitivity, and specificity. The Optimized Decision Tree (DT) algorithm achieved 73.54% accuracy using the Kaggle CVD dataset (12 attributes -70000 records). Saiful Islam [5] proposed some superior data analysis techniques such as Naive Bayes (NB), LR, DT. In this case, LR provided the highest accuracy with 86.25% on the UCI Heart Disease dataset (13 attributes, 302 records).

In 2019, Senthilkumar Mohan [6] proposed a method for improving prediction accuracy by defining key features and classifying them using a hybrid random forest with a linear model with 88.7% accuracy to predict heart disease. In this study they consider the UCI machine learning repository to collect the

dataset. In 2018, Mr. Chala Beyene [7] suggested some techniques for the occurrences of heart diseases using algorithms such as J48, Naive Bayes, and SVM. It achieved better accuracy around 89% on the StatLog heart disease dataset taken from the UCI machine learning laboratory that contains 13 attributes.

In [8], the author deals with the prediction of CVD by performing an analysis of six supervised machine learning algorithms, and feature selection was performed to find out the important risk factors. Some improvements in the quality of CVD prediction using a better preprocessing phase were proposed in [9].

Overall, many works were conducted for the prediction of CVD. Many datasets with different attributes had been considered by researchers to obtain better results and accuracy. Somehow, achieving better accuracy does depend on the selected attributes considered for the experiment. Our study considered two techniques and implemented them for the prediction of CVD. We have also considered important attributes required for the experiment.

III. DATASET

The experiments dataset was retrieved from the Kaggle data repository [10]. In this dataset, there are 70000 registers of patients and 12 attributes considered relevant for identifying the CVDs.

The dataset features were categorized into three types; the first type is objective (information as fact), the second type is examination (information from medical examination results), and the third type is subjective (information was given by the patient). The target variable is "cardio" which represents the existence or absence of CVD (binary value where yes=1, no =0). Table I. presents a description of each attribute.

TABLE I. DATASET FEATURES

Type	Name	Description
Objective	id	Patient's unique identifier int (unique identifier)
	age	Patient's age int (days)
	height	Patient's height int (cm)
	Weight	Patient's weight float (kg)
	Gender	Patient's gender int (1 = male, 2 = female)
Examination	ap_hi	Systolic blood pressure int(mmHg)
	ap_lo	Diastolic blood pressure int(mmHg)
	cholesterol	Patient's cholesterol int(1 = normal, 2 = above normal, 3 = well above normal)
	gluc	Patient's glucose int(1 = normal, 2 = above normal, 3 = well above normal)
Subjective	smoke	Patient smokes or not binary (yes=1, no =0)
	alco	Patient consumes alcohol or not binary (yes=1, no =0)
	active	Patient is physically active or not binary (yes=1, no =0)

IV. METHODOLOGY AND PROPOSED MODEL

The proposed model consist of the following steps: data preprocessing, features extraction, features selection, and CVD prediction algorithms (SVM, CNN).

A. *Data Preprocessing*

Data quality issues often need to be addressed. Therefore, the data cleaning process was applied. Specifically, the data was analyzed for the existence of duplicated data, missing values, outliers, and inconsistencies. In the dataset, there were no duplicated data, missing values, or inconsistencies, however, there are some outliers based on the following attributes: weight, height, ap_hi, and ap_lo.

Body mass index (BMI) was calculated to remove records that did not fit the standard BMI values (BMI < 150), and records that contain extremes values of blood pressure have been removed (average values for ap_hi [60:250], average values for ap_lo [10:200], and ap_hi > ap_lo).

Fig. 1. shows the balanced data distribution for the target attribute after data cleaning, where 49.46% of type "yes" (33986 registers) and 50.54% of type "no" (34725 registers).

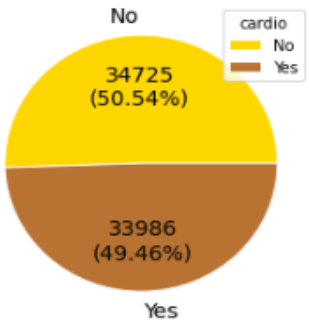

Fig. 1. Balanced data distribution.

B. *Features Extraction and Selection*

Table I. presents 12 features in the dataset. In addition, it is possible to extract three more features; body mass index (bmi), age in years rather than age in days (age_year) using the expression (age/365), and blood pressure different (ap_dif) between high and low pressures.

Feature selection has been successfully used in medical applications, where it can not only reduce dimensionality but also help us understand the causes of a disease. Correlation Matrix and Random Forest (RF) can be used as a feature selection technique [11].

Correlation matrix heatmap can identify which features are most related to the target variable. Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable). Fig. 2. shows the correlation matrix heatmap calculated from the dataset features and target attribute.

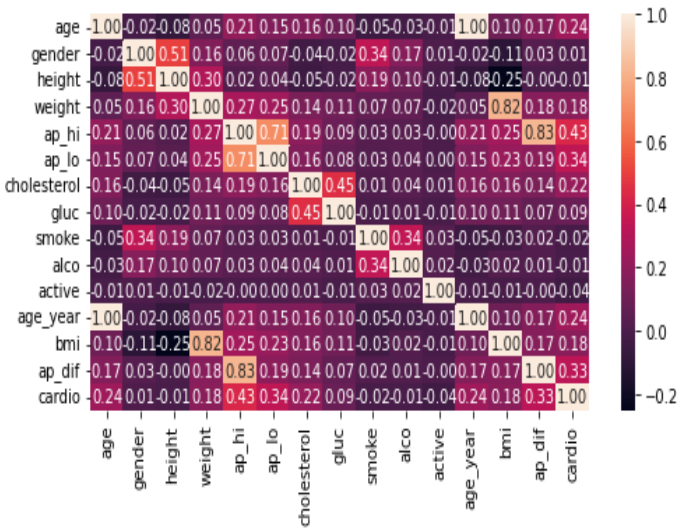

Fig. 2. Features correlation matrix heatmap

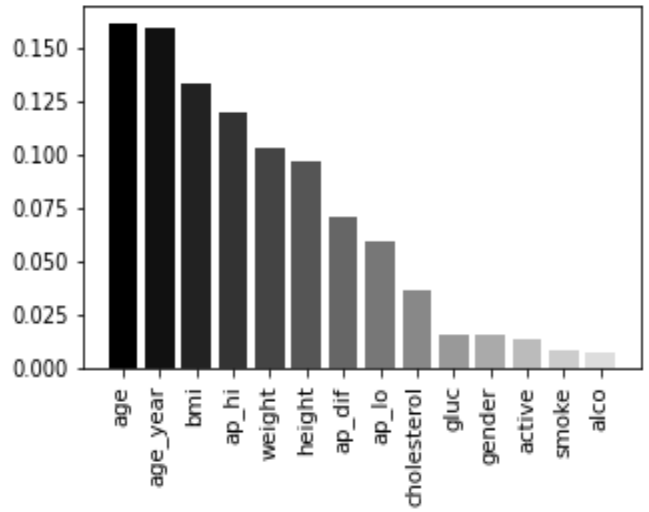

Fig. 3. Feature Importance based on RF

Random forest is a supervised learning algorithm for both classifications and regression. An important characteristic of random forests is that it can estimate the relative importance as part of the training process. The features importances are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. Fig. 3. shows the importance of each feature in the dataset.

The features { age_year, bmi, api_hi, weight, height, ap_dif, ap_lo, cholesterol } are selected to use during the training process, this selection is based on the results of both the correlation matrix heatmap(values > 0.1) and the RF importance of the features (values > 0.03). Features values are standardized by removing the mean and scaling to unit variance.

C. Support Vector Machine (SVM)

SVM is a supervised learning algorithm and binary classifier technique, which assumes that classes are linearly separable. SVM aims to find a hyperplane that gives the maximum margin between two classes of data.

In most cases, training data samples are not linearly separable, so this hyperplane does not exist. Using the kernel function gives an advantage to SVM for non-linearly separable classes[12]. SVM can be divided into two categories: support vector classification (SVC) and support vector regression (SVR).

SVC in scikit-learn was implemented based on the libsvm library. SVC was used during the training process on 80% of the dataset, with the Radial Basis function (RBF) as the kernel function.

D. Convolution Neural Network (CNN)

CNN is one of the most popular deep neural networks. It takes this name from the mathematical linear operation between matrixes called convolution. CNN has been designed to learn spatial hierarchies of features automatically [13].

CNN network model uses multiple building blocks, such as Conv1D layer, pooling layer, dropout layer, and flatten layer. The proposed CNN architecture (layers and activation functions) is shown in Fig. 4. The output shape is listed in Table II. for every layer in the proposed CNN architecture. The model was trained via minimizing the cross-entropy loss function through the Adam optimizer.

TABLE II. ARCHITECTURE LAYERS

Layer Type	**Output Shape** **(batch_shape, new_steps, filters)**
InputLayer	None,8,1
Conv1D	None,3,128
Dense	None,3,64
Dropout	None,3,64
Dense	None,3,32
Dropout	None,3,32
MaxPooling1D	None,1,32
Flatten	None,32
Dense	None,2

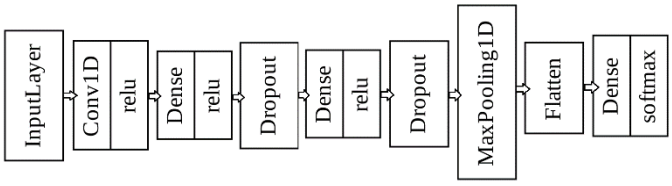

Fig. 4. CNN architecture

TABLE III. PERFORMANCE OF THE PROPOSED MODEL(%)

Model	A	P	R	S	F1	FA
SVM	73.25	75.84	66.27	79.84	70.74	20.16
CNN	73.98	75.98	68.84	77.62	71.82	22.38

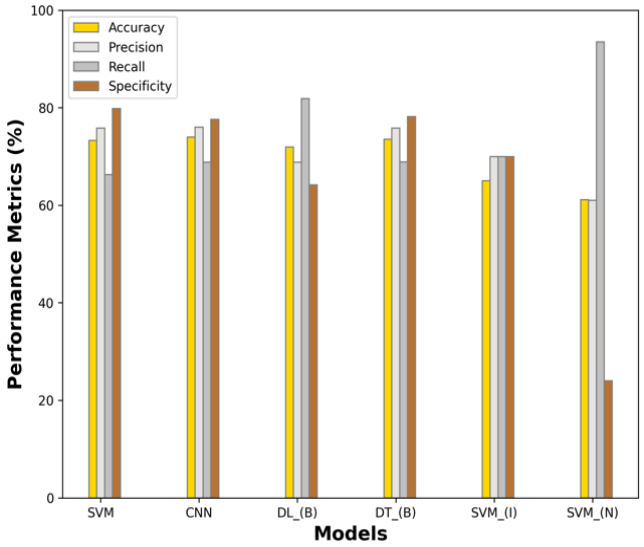

Fig. 5. Performance comparison of the proposed model and other benchmark models

V. EXPERIMENTAL RESULTS

Data preprocessing, features extraction, and classification algorithms (SVM, CNN) were implemented using the Google Colab platform (Linux-based hosted machine, Intel(R) Xeon(R) CPU @ 2.20GHz, and 13G of RAM) in Python using the Numpy, Scikit-learn, and Keras libraries, with Matplotlib for graph plotting [14][15].

Table III. shows the proposed model performance metrics; Accuracy (A), Precision (P), Recall (R), Specificity (S), F1 Score (F1), and False Alarm (FA). These metrics are calculated based on the confusion matrix.

DL_(B) and DT_(B) in [4] are Deep Learning (DL) and Decision Tree (DT), which use the same dataset and apply features selection based on importance without extracting new features.

SVM_(I) [8] used the same dataset and measured the effect of different attributes one by one, it used all features except two "alco", and "smoke". Our model used "alco" feature based on the importance of features. SVM_(N) [9] was applied in a different dataset (UCI dataset, all features without features selection).

The comparison between the proposed model and 4 different models is shown in Fig. 5. The proposed model obtains the highest performance metrics in accuracy and precision.

VI. CONCLUSION AND FUTURE WORK

In conclusion, it is possible to achieve scores over 70% to predict CVD using different data mining techniques e.g., SVM, and CNN. The results could be more indicative of CVD if different datasets with more features were used.

The number of false positives must be minimized as much as possible. So, there is still some future work to be done, such as applying ensemble techniques to build a more robust model with a better detection rate.

REFERENCES

[1] World Health Organization. "Fact Sheet: Cardiovascular diseases (CVDs) [Available from: https://www.who.int/en/news-room/fact-sheets /detail/cardiovascular-diseases-(cvds)", 2022.

[2] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar "Introduction to Data Mining (2nd Edition)", Pearson Education India, 2018.

[3] S. Verma and A. Gupta, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 249-253, doi: 10.1109/ICAIS50930.2021.9395963.

[4] Martins, B., Ferreira, D., Neto, C. *et al*. Data Mining for Cardiovascular Disease Prediction. *J Med Syst* **45,** 6 (2021). https://doi.org/10.1007/s10916-020-01682-8

[5] S. Islam, N. Jahan and M. E. Khatun, "Cardiovascular Disease Forecast using Machine Learning Paradigms," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 487-490, doi: 10.1109/ICCMC48092.2020.ICCMC-00091.

[6] Senthilkumar Mohan, Chandrasegar Thirumalai And Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", Special Section On Smart Caching, Communications, Computing and Cybersecurity For Information-centric Internet Of Things, IEEE Access Volume 7, 2019, DOI:10.1109/ACCESS.2019.2923707.

[7] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[8] I. A. Marbaniang, N. A. Choudhury and S. Moulik, "Cardiovascular Disease (CVD) Prediction using Machine Learning Algorithms," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342297.

[9] N. Louridi, M. Amar and B. E. Ouahidi, "Identification of Cardiovascular Diseases Using Machine Learning," 2019 7th Mediterranean Congress of Telecommunications (CMT), 2019, pp. 1-6, doi:10.1109/CMT.2019.8931411.

[10] S. Ulianova, "Cardiovascular disease dataset, version 1," 2019. Accessed: Jan. 25, 2022. [Online]. Available: https://www.kaggle.com/sulianova /cardiovascular-disease-dataset

[11] Rashme, Tamanna Yesmin, et al. "Early Prediction of Cardiovascular Diseases Using Feature Selection and Machine Learning Techniques." 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2021.

[12] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM transactions on intelligent systems and technology (TIST) 2, no. 3 (2011): 1-27.

[13] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186

[14] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[15] Bisong, Ekaba. "Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners". Apress, 2019